On glottal source shape parameter transformation using a novel deterministic and stochastic speech analysis and synthesis system

Stefan Huber[1]*, Axel Roebel*[2]

[1],[2] Sound Analysis/Synthesis Team, IRCAM-CNRS-UPMC STMS, 75004 Paris, France

stefan (dot) huber, axel (dot) roebel (at) ircam (dot) fr

Abstract

In this paper we present a flexible deterministic plus stochastic model (DSM) approach for parametric speech analysis and synthesis with high quality. The novelty of the proposed speech processing system lies in its extended means to estimate the unvoiced stochastic component and to robustly handle the transformation of the glottal excitation source. It is therefore well suited as speech system within the context of Voice Transformation and Voice Conversion. The system is evaluated in the context of a voice quality transformation on natural human speech. The voice quality of a speech phrase is altered by means of resynthesizing the deterministic component with different pulse shapes of the glottal excitation source. A subjective listening test suggests that the speech processing system is able to successfully synthesize and arise to a listener the perceptual sensation of different voice quality characteristics. Additionally, improvements of the speech synthesis quality compared to a baseline method are demonstrated.

Index Terms: Parametric speech analysis / synthesis, Glottal source, Voice quality, LF model, R_d shape parameter

1. Introduction

In this paper we present a method to transform the deterministic part of the glottal excitation source. The main motivation of the following paper is the presentation of an improved method for coherent modification of the glottal pulse shape. The glottal pulse shape is generally accepted to reflect different phonation types of human voice production [1] and one of the important parameters determining the perceived voice quality that is strongly related to the vocal effort [2, 3]. The terminology used in the following is describing the lax-tense dimension of the voice quality [2, 4] distinguishing tense (pressed), modal (normal), and relaxed (breathy) voice qualities.

Recent research in the speech community has notably improved the speech synthesis quality by explicitly modelling the deterministic and stochastic component of the glottal excitation source [5, 6]. Advanced source-filter decomposition strategies as in [7, 8, 9] address finer details defined by extended voice production models for human speech. These approaches analyze an extended voice descriptor set to model their transformation and synthesis. The extended voice descriptor set consists of: the Vocal Tract Filter (VTF), the glottal pulse positions and shapes, and energies and a random component described by spectral and temporal envelopes.

In this paper we present a novel speech analysis and synthesis system based on [9]. The proposed system is a Deterministic plus Stochastic Model (DSM). It extracts the unvoiced stochastic component from a speech signal by subtracting the corresponding voiced deterministic component [10]. The proposed system separately models the stochastic and deterministic components and does therefore not correspond to the classical source and filter model. The contribution of the following research and the advancements compared to the baseline method lies in the extended means to estimate the unvoiced stochastic component, to robustly extract the VTF and to handle the variations in energy and signal behaviour implied with glottal source transformations.

The paper is organized as follows. Section 2 presents the novel speech framework. Section 3 discusses the aspects of voice quality transformation. Section 4 introduces a state-of-the-art speech processing system. Section 5 presents a subjective evaluation based on a listening test of natural human speech. Section 6 concludes with the findings studied in this paper.

Initial work on this problem was financed by a CIFRE contract between IRCAM and the Acapela Group. A part of the investigation was funded by the ANR Project ChaNTeR supporting voice quality conversion for singing synthesis.

2. DSM-based Parametric Re-Synthesis

The proposed speech analysis and synthesis system is designed for the utilization as a basic component in the context of advanced voice transformation applications. It is denoted **PaReSy** for **Pa**rametric speech analysis **Re**-**Sy**nthesis.

2.1. Voice production model

PaReSy operates upon the following generic interpretation of the human voice production

$$s(n) = u(n) + v(n) = u(n) + \sum_i g(n, P_i) * c(n, P_i) * \delta(n - P_i) \quad (1)$$

Here $s(n)$ is the speech signal that is represented by means of a stochastic (unvoiced) component $u(n)$ and a deterministic (voiced) component $v(n)$. The deterministic component contains the sequence of glottal pulses that are located at the time positions P_i that each represent a Glottal Closure Instant (GCI) with index i. Each glottal pulse is represented in terms of the glottal flow derivative $g(n, P_i)$ and is convolved with the Vocal Tract Filter (VTF) that is active for the related position $c(n, P_i)$ and a Dirac impulse at the GCI P_i.

Using this model we make the following assumptions: The VTF $c(n, P_i)$ is supposed to be minimum phase [11]. The glottal pulse derivative $g(n, P_i)$ is used to represent the glottal pulse using the Liljencrants-Fant (LF) model, and the effect of the lip radiation [12]. The LF model is parameterized by a scalar shape parameter R_d [13, 14], which is estimated as described in [15, 16]. Changing R_d continuously from lower to higher values will allow changing the LF pulse shape on a continuum from tense to relaxed voice qualities.

For being able to make spectral domain manipulations the speech signal model given in equ. (1) is processed in the spectral domain using the short time Fourier transform (STFT). An efficient STFT representation requires further approximations.

The sliding (Hanning) window $w_h(n)$ that is used to calculate the STFT selects a signal segment covering a few consecutive glottal pulses $g(n, P_i)$ each related to a slightly different VTF. Here we will assume that both the glottal pulse shape and the VTF do not change within the window and are given approximately by the corresponding parameters in the window center. We further assume that the filtering processes implied by each convolutional operation between the signal components of equ. (1) is involving impulse responses that are shorter than the window length. The STFT of the speech signal is then given by

$$\begin{aligned} S(\omega, m) &= U(\omega, m) + V(\omega, m) && (2) \\ &= (U(\omega, m) + G(\omega, m)C(\omega, m)H(\omega, m)) && (3) \end{aligned}$$

Here m is the position of the window center and ω the frequency variable of the discrete time Fourier transform (FT). For brevity the dependency of all signal spectra with respect to m will be dropped in the follwoing. $U(\omega)$ and $V(\omega)$ are the FT of the windowed voiced and unvoiced signals from equ. (1) under the assumption that g and c and the corresponding FT spectra $G(\omega)$ and $C(\omega)$ are quasi stationary within the window. The radiation filter at lips and nostrils level $R(\omega)$ is not explicitly present in the PaReSy model, but implicitly represented in the glottal flow derivative $G(\omega)$ and the unvoiced component $U(\omega)$.

2.2. Glottal source synthesis and VTF extraction

The glottal shape parameter R_d is estimated using the best phase minimization variant proposed in [15]. It constructs the error lattice for the Viterbi smoothing proposed in [16]. The resulting R_d estimation is calculated on the STFT time grid but assigned to the closest GCI which are derived using the method described in [17]. The spectral envelope sequence $\mathcal{T}_{sig}$ is estimated on the input signal $s(n)$ using the True Envelope estimator described in [18]. Another spectral envelope sequence $\mathcal{T}_g$ is estimated on the synthesized glottal pulse derivative sequence $\sum_i g(n, P_i) * \delta(n - P_i)$. The extraction of the vocal tract filter $C(\omega)$ is obtained by means of the full-band division of $\mathcal{T}_{sig}$ by $\mathcal{T}_g$. The utilization of $\mathcal{T}_g$ in the division is required to suppress the spectral ripples occurring for higher R_d values [14, 19].

2.3. Estimation of the unvoiced stochastic component

The separation of a speech signal into the contributions of the voiced deterministic $V(\omega)$ and the unvoiced stochastic component $U(\omega)$ is based on the calculation of a residual of a sinusoidal model. Using the sinusoidal model has the advantage that pulse shape errors that are due to the rather limited coverage of the R_d parameterization of the LF model will not lead to an increase in the unvoiced component. The following two algorithmic steps a) and b) below present a means to robustly extract the unvoiced component $U_{res}(n)$ of the signal $s(n)$.

a) Remix with demodulation: This approach aims to simplify the sinusoidal detection by de-modulating the F_0 contour and the Hilbert amplitude envelope $\mathcal{H}$ from the signal $s(n)$. First the original F_0 contour of $s(n)$ is warped to become flat by means of time varying re-sampling using as target F_0 the mean of the original fundamental frequency contour. The resampling operation will locally and globally change the time duration of all signal features, which however is not a problem because the effect can be inverted after the extraction of the residual. The varying amplitude contour of $s(n)$ is demodulated by means of dividing the signal by its smoothed Hilbert transform $\mathcal{H}(s(n))$ similar as in [22, 5]. Here however, the smoothing kernel is simply the Hanning window of duration exactly equal to $4/F_T$ which will optimally remove all envelope fluctuations that are related to the deterministic components. The resulting signal $s_{flat}(n)$ is flat in amplitude envelope and fundamental frequency facilitating the detection of sinusoids following [20] even for relatively high harmonic numbers avoiding energy shift between voiced and unvoiced components [21]. The sinusoidal content is subtracted from $s_{flat}(n)$ and the demodulation steps are inverted so that the the original AM-FM modulation is recreated. This generates the unvoiced residual signal $u_{res}(n)$.

b) Scale to $\mathcal{T}_{sig}$ level and noise excitation: The sinusoidal detection of step a) may be erroneous for some signal segments such as fast transients. The scaling described in equ. 4 minimizes the difference between the unvoiced stochastic signal spectrum $U_{res}(\omega)$ and the observed signal spectrum $S(\omega)$ above the Voiced / Unvoiced Frequency boundary ω_{VU} [23] up to the Nyquist frequency ω_{nyq}:

$$\begin{aligned} \eta &= \frac{1}{\omega_{nyq} - \omega_{VU}} \int_{\omega_{VU}}^{\omega_{nyq}} \left(\mathcal{T}_{sig}^{dB}(\omega) \text{-} \mathcal{T}_{unv}^{dB}(\omega) \right) d\omega \\ \mathcal{T}_{unv}(\omega) &= \mathcal{T}_{unv}(\omega)(1\text{-}\omega_{VU}/\omega_{nyq}) \cdot 10^{\eta/20}. \end{aligned} \quad (4)$$

Here the dependency with m has been neglected. η equals the mean difference in dB between $\mathcal{T}_{sig}$ and the spectral envelope $\mathcal{T}_{unv}$ estimated on $U(\omega)$. The scaling of $\mathcal{T}_{unv}$ is additionally weighted by the time-varying ratio of F_{VU} versus F_{nyq}. The multiplication of the STFT of a white noise signal with the envelope $\mathcal{T}_{unv}(\omega)$ generates the unvoiced signal STFT $U(\omega)$.

2.4. Energy modelling

A simple Root-Mean-Square (RMS) measure F_{RMS} evaluates the effective energy value E on the linear amplitude spectrum A_{lin}=$|Y(\omega)|$ of any arbitrary signal spectrum $Y(\omega)$. The RMS energy measures are estimated in PaReSy as defined in equ. (5):

$$\begin{aligned} F_{RMS}(A_{lin}, k) &= \sqrt{1/K \cdot \Sigma^K \left(A_{lin}(k)^2\right)} \\ E_{sig} &= F_{RMS}(|S(\omega)(t)|) \\ E_{unv} &= F_{RMS}(|U(\omega)(t)|) \\ E_{voi} &= E_{sig} - E_{unv} \end{aligned} \quad (5)$$

E_{sig} and E_{unv} measure the RMS energy of signal $S(\omega)$ and the unvoiced component $U(\omega)$. The energy E_{voi} of the voiced component $V(\omega)$ is expressed as their difference. A transformed $R_d^{'}$ contour causes an altered energy value E'_{voi} measured on the transformed voiced component $V'(\omega)$, with the operator $'$ indicating a transformation. The high (low) pass filtering applied to $U(\omega)$ $(V(\omega))$ explained in section 2.6 generates as well an energy change. A re-scaling of the energy to the original energy measures ensures their maintenance.

2.5. GMM-based F_{VU} prediction

The spectral fading synthesis presented in the following section 2.6 requires a transformed F'_{VU} frequency value. F'_{VU} is predicted using a modified GMM approach detailed in [24, 25, 16]. The GMM model $\mathcal{M}$ is trained on the voice descriptor set d=[R_d,F_0, $H1$-$H2$,E_{voi}, E_{unv}] and the F_{VU} reference value r. The descriptors of d are chosen due to their high correlation with r. $H1$-$H2$ refers to the amplitude difference in dB of the first two harmonic sinusoidal partials. The prediction function

$$F(d) = \sum_{q=1}^{Q} p_q^d(d) \cdot [\mu_q^r + \Sigma_q^{rd} \Sigma_q^{dd^{-1}} (d - \mu_q^d)] \quad (6)$$

is derived from $\mathcal{M}$ by the definition of equ. 6, with Q=15 being the number of utilized Gaussian mixture components. An initial F_{VU}^p value prediction is computed from $F(d)$. An error GMM model $\mathcal{M}_{err}$ is trained on the modelling error

$$\epsilon_M = \sqrt[2]{(F_{VU} - F_{VU}^p)^2} \quad (7)$$

serving as reference value r_ϵ=ϵ_M, and on the voice descriptor set d. The transformed descriptor counterpart d' contains the original F_0 contour but transformed values for the remaining voice descriptors: d'=[R'_d, F_0, $H'1$-$H'2$, E'_{voi}, E'_{unv}]. The

GMM-based modelling to predict a F'_{VU} contour from the voice descriptor sets d and d' is defined by the following equations:

$$F^{p}_{VU_\mu} = \mathcal{M}\left(F(d)\right) \quad (8)$$

$$F^{'p}_{VU_\mu} = \mathcal{M}\left(F(d')\right) \quad (9)$$

$$F^{p}_{VU_\sigma} = \mathcal{M}_{err}\left(F_{err}(d)\right) \quad (10)$$

$$F^{'p}_{VU_\sigma} = \mathcal{M}_{err}\left(F_{err}(d')\right) \quad (11)$$

$$F'_{VU} = F^{'p}_{VU_\mu} + (F_{VU} - F^{p}_{VU_\mu}) \cdot F^{'p}_{VU_\sigma} / F^{p}_{VU_\sigma} \quad (12)$$

Each trained model pair $\mathcal{M}$ and $\mathcal{M}_{err}$ is utilized to predict via their derived prediction functions F and F_{err} the mean prediction value $F^{p}_{VU_\mu}$ and the predicted standard deviation $F^{p}_{VU_\sigma}$ from descriptor set d, and likewise for the transformed set d'. The "true" prediction value would equal $F^{p}_{VU_\mu}$ if no model error occurs: ϵ_M=0. The calculation of F'_{VU} from the transformed d' and the original voice descriptor set d is defined by equ. 12. It evaluates the difference between the original F_{VU} and the predicted $F^{p}_{VU_\mu}$ value. The difference result is normalized by the ratio of the original and transformed standard deviations $F^{'p}_{VU_\sigma}$ and $F^{p}_{VU_\sigma}$ of the modelled data distribution, and corrected by the transformed predicted mean value $F^{'p}_{VU_\mu}$.

2.6. Spectral fading synthesis

The PaReSy synthesis variant "Spectral fading" is designed to handle voice quality transformations by suppressing possibly occurring artefacts. Here a short summary discusses the impact of R_d on the spectral slope required to understand the motivation for the spectral fading synthesis presented in this section. The glottal source shape parameter R_d is strongly correlated with the spectral slope. R_d changes lead to changes of the spectral slope. References to an extensive analysis of the spectral correlates of R_d can be found in [13, 14, 26, 27, 19]. A more relaxed voice quality is reflected by higher R_d values and is related to a sinusoidal-like glottal flow derivative which generates higher spectral slopes. A more tense voice quality is parameterized by lower R_d values and related to an impulse-like glottal flow derivative which produces lower spectral slopes. A lower (higher) spectral slope indicates that more (less) sinusoidal content can be observed in higher frequency regions. The voice quality transformation to change an original speech recording having a modal voice quality to a more tense voice character has to extend the quasi-harmonic sequence of sinusoidals above the F_{VU}. Contrariwise, a transformation to a more relaxed voice quality needs to reduce the sinusoidal content. A modification of the glottal excitation source required for voice quality transformations implies thus a F_{VU} modification. The altered F'_{VU} frequency has to be naturally represented by properly joining the voiced $V(\omega)$ and unvoiced $U(\omega)$ signal components. The transformation of the original R_d contour used to extract $C(\omega)$ introduces an energy variation in the re-synthesis of a transformed $V'(\omega)$. However, even with the energy maintenance of section 2.4 the alteration of a modal to a very tense voice quality may result into sinusoidal content being of higher energy than the noise part at F_{nyq}. This sets $F'_{VU} = F_{nyq}$ and causes audible artefacts. Therefore F'_{VU} is predicted using the method described in section 2.5. Additionally, the spectral fading method employs two spectral filters to cross fade $V(\omega)$ and $U(\omega)$ at the F'_{VU} frequency. The spectral band around F_{VU} is comprised of a mix of both voiced deterministic $V(\omega)$ and unvoiced stochastic $U(\omega)$ components. A low pass filter P_L fades out the voiced component $V(\omega)$ and a high pass filter P_H fades in the unvoiced component $U(\omega)$ with increasing frequency. The linear ramps with a slope of m_{LP}=-96 dB and m_{HP}=-48 dB per octave define the steepness of the low pass P_L and respectively the high pass P_H filter. A higher value is chosen for m_{LP} since the F'_{VU} prediction may be very high for very tense voice qualities. A less steep fade out filter would not be effective enough to suppress artefacts.

3. Voice quality transformation

The study of [28] on the Just Noticable Differences (JND) of human auditory perception reports that changes in higher (lower) value regions of Open Quotient OQ (asymmetry coefficient α_m) require longer distances of ΔOQ ($\Delta\alpha_m$) to arise the sensation of a voice quality change in the perception of a listener. We spread according to that hypothesis the original R_d contour into several R_d contours with positive and negative offsets covering the complete R_d range such that lower ΔR_d steps are placed in lower and higher ΔR_d steps in higher R_d value regions. One example is illustrated in fig. (1) on the phrase em-

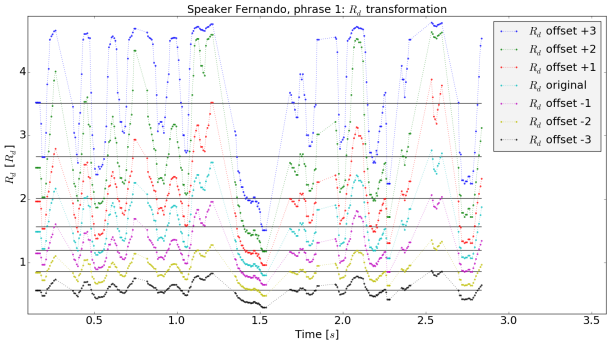

Figure 1: *Generated R_d contour examples*

ployed for the evaluation in section 5. Table (1) shows the mean R^{μ}_{d} values of the original R_d contour with index 0, and respectively 3 positive and 3 negative μ values for each voice quality change. $R^{\sigma^2}_{d}$ lists their variance σ^2. It increases with increasing R_d to reflect the hypothesis of having to apply higher ΔR_d steps with higher R_d values. The R^{μ}_{d} (diff) column reflects the mean ΔR_d steps measured between each row index on the R^{μ}_{d} values. As well the μ difference increases with increasing R^{μ}_{d}.

Table 1: *R_d value example for voice quality transformation*

Voice quality (index)	R^{μ}_{d}	$R^{\sigma^2}_{d}$	R^{μ}_{d} (diff)
Very relaxed (+3)	3.5109	0.9031	-0.8397
Relaxed (+2)	2.6711	0.7825	-0.6597
Modal to relaxed (+1)	2.0114	0.3631	-0.4442
Modal (original) (0)	1.5673	0.1937	
Tense to modal (-1)	1.1936	0.0941	-0.3737
Tense (-2)	0.8601	0.0341	-0.3335
Very tense (-3)	0.5704	0.0154	-0.2898

4. Baseline method SVLN

The method called "**S**eparation of the **V**ocal tract with the **L**iljencrants-Fant model plus **N**oise" detailed in [29, 30, 9] represents the baseline method on whose means the proposed system PaReSy is build upon. The main difference lies in the VTF representation, the energy model and the estimation of the stochastic noise component. SVLN constructs the latter by high pass filtering white noise, applying an amplitude modulation parameterized by the glottal pulse sequence, and cross fading between consecutive synthesized noise segments. The gain σ_g measures the energy level at F_{VU} at analysis to control the stochastic energy at the synthesis step. SVLN synthesizes glottal pulses with the LF model in the spectral domain to extract $C(\omega)$ below F_{VU}. The VTF above F_{VU} is taken from the

signals spectral envelope. SVLN facilitates voice quality transformations while maintaining a high synthesis quality [31, 9].

5. Evaluation

This sections presents the results of a listening test conducted on natural human speech of French speaker "Fernando" having an Hispanic accent. The baseline method SVLN of section 4 and the proposed method PaReSy of section 2 received the same voice descriptors R_d, F_0 and F_{VU} as pre-estimated input to analyze $C(\omega)$. Please note that SVLN requires to smooth the voice descriptor contours. Due to the energy measure at F_{VU} it cannot handle value changes varying too quickly in short-time segments [29]. For this test a median smoothing filter covering 100 ms was applied. The PaReSy spectral fading synthesis variant presented in 2.6 requires the F_{VU} prediction of section 2.5. An example is depicted in fig. (2). The transformed $R_d^{'}$

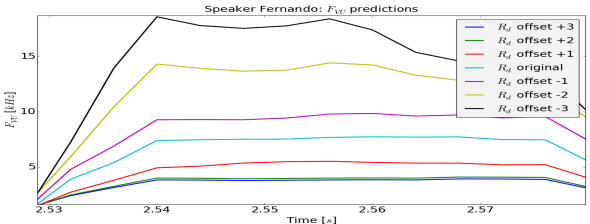

Figure 2: *PaReSy F_{VU} prediction excerpt*

contours and the original R_d contour were employed by both systems for synthesis. Following the voice production model of equ. (3), a transformed glottal pulse $G^{'}_{R_d}(\omega)$ leads to a transformed reconstructed signal $S'(\omega)$. The unvoiced component $U(\omega)$ remains unmodified. 11 participants rated each speech phrase according to the voice quality characteristics given in the first column of table (1). The voice quality assessment examines how well both synthesis systems are able to produce different voice quality characteristics. A second evaluation metric examines the synthesis quality on the Mean Opinion Score (MOS) scale. Fig. (3) depicts the voice quality ratings for the

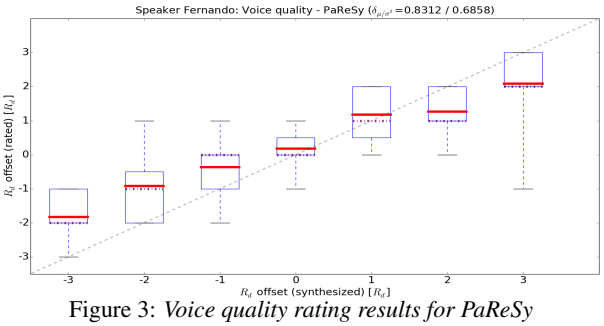

Figure 3: *Voice quality rating results for PaReSy*

proposed method PaReSy. The small horizontal grey lines at both ends (whiskers) are set to show the minimum and maximum value for each evaluation. The horizontal red (violet) lines reflect the mean (median) voice quality ratings of all participants per test phrase with the same indices as in table (1). The diagonal grey dashed line exemplifies their ideal placement if each test participant would have been able to associate perceptually each synthesized voice quality example to its corresponding voice quality characteristic. The mean deviation value $\delta_\mu = 0.83$ expresses the disagreement of the listeners, being ideally $\delta_\mu = 0.0$. A higher mean deviation value δ_μ=0.97 as compared to PaReSy indicates for the baseline method SVLN shown in fig. (4) that the listeners could less well capture the different synthesized voice qualities. Clear voice quality associations can be concluded for both systems. Both follow

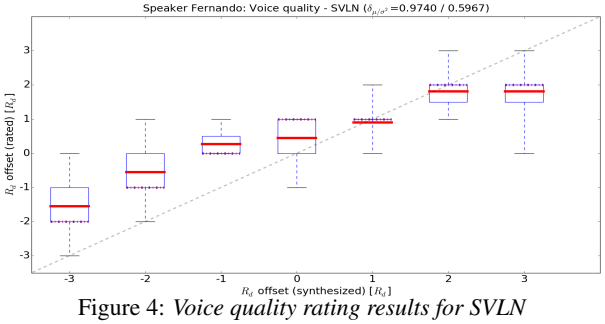

Figure 4: *Voice quality rating results for SVLN*

roughly the ideal dashed grey line with the deviations increasing with higher changes. The MOS synthesis quality evaluation

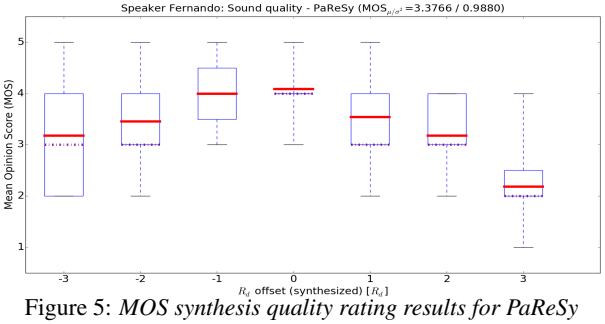

Figure 5: *MOS synthesis quality rating results for PaReSy*

for PaReSy shown in fig. (5) exhibits partially highest ratings up to an excellent synthesis quality of 5 for all but the "relaxed" and "very relaxed" voice quality characteristics with index +2 and +3. The evaluated mean synthesis quality MOS$_\mu$=2.82 of

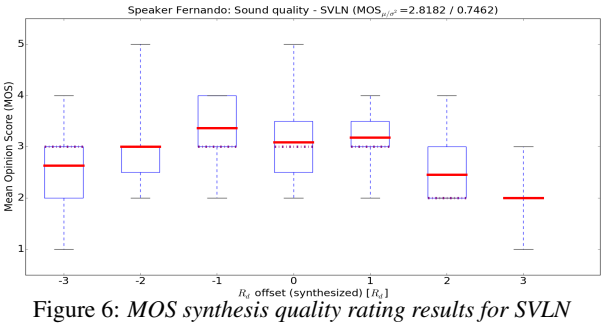

Figure 6: *MOS synthesis quality rating results for SVLN*

SVLN shown in fig. (6) is comparably lower than MOS$_\mu$=3.38 for PaReSy. Stronger voice quality changes are assessed with less good MOS synthesis qualities for both systems. In general, PaReSy received a lower deviation from the true voice quality rating and a higher MOS synthesis quality compared to SVLN, shown in table (2).

Table 2: *Voice quality (VQ) and MOS sound quality ratings*

Method	Δ VQ$_\mu$	Δ VQ$^2_\sigma$	MOS$_\mu$	MOS$_\sigma$
PaReSy	0.8312	0.6858	3.3766	0.9880
SVLN	0.9740	0.5967	2.8182	0.7462

6. Conclusions

The findings presented with the subjective listening test of section 5 suggest that the proposed novel speech analysis and synthesis system is able to analyze an input speech phrase such that different re-synthesized versions carry the perception of different voice quality characteristics. Its assessed synthesis quality received partially very good judgements for minor changes in voice quality. Major voice quality changes are appraised of moderate quality for both the baseline and the proposed method.

7. References

[1] D. G. Childers and C. K. Lee, "Vocal quality factors: analysis, synthesis, and perception." *Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–410, 1991.

[2] J. Kane, S. Scherer, M. Aylett, L.-P. Morency, and C. Gobl, "Speaker and language independent voice quality classification applied to unlabeled corpora of expressive speech," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[3] J.-S. Liénard and C. Barras, "Fine-grain voice strength estimation from vowel spectral cues," in *14th Annual Conference of the Int. Speech Communication Association (Interspeech ISCA)*, Lyon, France, 2013, pp. 128–132.

[4] J. D. M. Laver, *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980, vol. 31.

[5] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 968–981, 2012.

[6] J. P. Cabral and J. Carson-Berndsen, "Towards a better representation of the envelope modulation of aspiration noise," in *Advances in Nonlinear Speech Processing*, ser. Lecture Notes in Computer Science, T. Drugman and T. Dutoit, Eds. Springer Berlin Heidelberg, 2013, vol. 7911, pp. 67–74.

[7] D. Vincent, O. Rosec, and T. Chonavel, "A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and hnm modeling," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, 2007, p. 525–528.

[8] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *9th Annual Conference of the Int. Speech Communication Association (Interspeech ISCA)*, Brisbane, Australia, September 2008, pp. 1829–1832.

[9] G. Degottex, P. Lanchantin, A. Röbel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, 2013.

[10] X. Serra, *Musical Sound Modeling with Sinusoids plus Noise*. Swets and Zeitlinger, 1997, pp. 91–122. [Online]. Available: files/publications/MSM-1997-Xserra.pdf

[11] R. Maia and Y. Stylianou, "Complex cepstrum factorization for statistical parametric synthesis," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2014, pp. 3839–3843.

[12] G. Fant, J. Liljencrants, and Q.-G. Lin, "A four-parameter model of glottal flow," *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, vol. 26, no. 4, pp. 001–013, 1985.

[13] G. Fant, "The lf-model revisited. transformation and frequency domain analysis," *Quaterly Progress and Status Report, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*, vol. 36, no. 2-3, pp. 119–156, 1995.

[14] ——, "The voice source in connected speech," *Speech Communication*, vol. 22, no. 2-3, pp. 125–139, 1997.

[15] S. Huber, A. Röbel, and G. Degottex, "Glottal source shape parameter estimation using phase minimization variants," in *13th Annual Conference of the Int. Speech Communication Association (Interspeech ISCA)*, ser. 1990-9772, Portland, Oregon, USA, 2012, pp. 1644–1647.

[16] S. Huber and A. Röbel, "On the use of voice descriptors for glottal source shape parameter estimation," *Computer Speech & Language*, vol. 28, no. 5, pp. 1170 – 1194, 2014.

[17] G. Degottex, A. Röbel, and X. Rodet, "Joint estimate of shape and time-synchronization of a glottal source model by phase flatness," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA, 2010, pp. 5058–5061.

[18] A. Röbel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modelling with unknown model order," *Elsevier, Pattern Recognition Letters*, vol. 28, no. 11, pp. 1343 – 1350, 2007.

[19] C. d'Alessandro, B. Bozkurt, B. Doval, T. Dutoit, N. Henrich, V. Tuan, and N. Sturmel, "Phase-based methods for voice source analysis." in *Advances in Nonlinear Speech Processing*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2007, vol. 4885, pp. 1–27.

[20] M. Zivanovic and A. Röbel, "Adaptive threshold determination for spectral peak classification," *Computer Music Journal*, vol. 32, no. 2, pp. 57–67, 2008.

[21] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 12–23, 1998.

[22] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 4609–4612.

[23] T. Drugman and Y. Stylianou, "Maximum voiced frequency estimation: Exploiting amplitude and phase spectra," *Signal Processing Letters, IEEE*, vol. 21, no. 10, pp. 1230–1234, Oct 2014.

[24] P. Lanchantin and X. Rodet, "Dynamic model selection for spectral voice conversion," in *11th Annual Conference of the Int. Speech Communication Association (Interspeech ISCA)*, Makuhari, Chiba, Japan, 2010, pp. 1720–1723.

[25] ——, "Objective evaluation of the dynamic model selection method for spectral voice conversion," in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 5132–5135.

[26] B. Doval and C. d'Alessandro, "The spectrum of glottal flow models," Laboratoire d'informatique pour la mécanique et les sciences de l'ingénieur (Orsay), Orsay, Tech. Rep. LIMSI 99-07, 1999.

[27] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[28] N. Henrich, C. d'Alessandro, B. Doval, M. Castellengo, G. Sundin, and D.Ambroise, "Just noticeable differences of open quotient and asymmetry coefficient in singing voice," *Journal of Voice*, vol. 17, no. 4, pp. 481–494, 2003.

[29] G. Degottex, "Glottal source and vocal tract separation," Ph.D. dissertation, Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Université de Pierre et Marie Curie (UPMC), Université de Paris 6, Paris, France, 2010.

[30] G. Degottex, A. Röbel, and X. Rodet, "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5128–5131.

[31] P. Lanchantin, G. Degottex, and X. Rodet, "A hmm-based speech synthesis system using a new glottal source and vocal-tract separation method," in *Proc. IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA, 2010, pp. 4630–4633.